TRANSIENT INTERFERENCE SUPPRESSION IN SPEECH SIGNALS BASED ON THE OM-LSA ALGORITHM

Ariel Hirszhorn, David Dov, Ronen Talmon and Israel Cohen

Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
{arielhir@tx, davidd@tx, ronenta2@tx, icohen@ee}.technion.ac.il

ABSTRACT

Typical transient interferences, e.g. door knocks and keyboard tapping, are short in time, widely spread across the frequency domain, and have an abrupt nature. Thus, traditional speech enhancement techniques that use temporal smoothing to estimate the power spectral density (PSD) of the interference are inadequate. In this paper, we present a speech enhancement algorithm that suppresses transient interferences and pseudo-stationary background noise. The algorithm comprises an estimation of the transient and the pseudo-stationary noise PSDs, and enhancement of speech. The proposed algorithm is capable of tracking rapid variations of the input signal spectra and enables to effectively estimate the PSD of the transients. Experimental results show that the proposed algorithm is robust, does not rely on transient periodicity or reoccurrence, and exhibits good performance for various transient interference types.

Index Terms— Speech enhancement, speech processing, acoustic noise, impulse noise, transient noise.

1. INTRODUCTION

Transients are undesired abrupt interferences which can be originated by keyboard typing, construction operations, knocking, hammering, etc. Characteristic transients are featured by a short time duration and a wide spread over the frequency domain with respect to speech phonemes. Traditional speech enhancement techniques assume pseudo-stationary noise, which enables to estimate the power spectral density (PSD) of the noise via temporal smoothing [1, 2, 3, 4]. This assumption does not hold for transient interferences due to their fast varying nature.

Recently, an approach based on nonlocal filtering has been proposed to enhance speech interfered by transients [5, 6, 7]. First, the transient interferences are enhanced using a modified speech estimator. Then, diffusion maps [8] is used to learn the geometric structure of the transients, which is in turn utilized to estimate the transients PSD using nonlocal diffusion filtering. Finally, the speech is enhanced and the interferences are suppressed by the optimally-modified LSA (OM-LSA) filter equipped with an estimate of the transients PSD. Unfortunately, the main drawback imposed by the nonlocal diffusion filtering is the key assumption that the same interference pattern appears several times in the measurement. Thus, a single transient is generally not identified as interference, and hence not suppressed.

This research was supported by the Israel Science Foundation (grant no. 1130/11).

In this paper, we present an algorithm that enhances speech corrupted by transient interferences and pseudo-stationary noise. The algorithm comprises an estimation of the transient and the pseudo-stationary noise PSDs, and enhancement of speech. We introduce a modified version of [6] based entirely on extensions of the OM-LSA algorithm [4]. The proposed algorithm is capable of tracking rapid variations of the input signal spectra and enables to efficiently estimate the PSD of the transients. We show that the proposed solution is robust to the type of transient noise, does not require off-line nor pre- or post-processing, and does not rely on transient periodicity or recurrence.

The remainder of this paper is structured as follows. In Section 2, we formulate the problem. In Section 3, we present the proposed algorithm, and in Section 4, experimental results demonstrate the performance for various transient interferences.

2. PROBLEM FORMULATION

Let $x(n)$ denote a speech signal and let $d(n)$ and $t(n)$ denote an additive stationary (or quasi-stationary) noise and a transient interference, respectively. The measured signal $y(n)$ is given by:

$$y(n) = x(n) + t(n) + d(n). \tag{1}$$

Let $Y(k,l)$ be the measured signal represented in the time-frequency domain by applying the short-time Fourier transform (STFT):

$$Y(k,l) = \sum_{n=0}^{N-1} y(n+lM)h(n)e^{-j\frac{2\pi}{N}nk} \tag{2}$$

where k is the frequency bin index, l is the time frame index, h is an analysis window (e.g. Hamming window) of length N, and M is the number of overlapping samples between two successive frames. Applying the STFT on each component in (1) yields

$$Y(k,l) = X(k,l) + T(k,l) + D(k,l) \tag{3}$$

where $X(k,l)$, $T(k,l)$ and $D(k,l)$ are the STFTs of $x(n)$, $t(n)$ and $d(n)$, respectively. The objective is to find an estimator $\widehat{X}(k,l)$ for the speech signal from the measured signal $Y(k,l)$ for each spectral component.

Let $\lambda_y(k,l) = \mathbb{E}\left[|Y(k,l)|^2\right]$ be the spectral variance of the measured signal. We assume that the speech, the transient interference, and the stationary noise are uncorrelated. Thus, the spectral variance of the measurement is given by

$$\lambda_y(k,l) = \lambda_x(k,l) + \lambda_t(k,l) + \lambda_d(k,l) \tag{4}$$

where $\lambda_x(k,l) = \mathbb{E}\left[|X(k,l)|^2\right]$, $\lambda_t(k,l) = \mathbb{E}\left[|T(k,l)|^2\right]$, and $\lambda_d(k,l) = \mathbb{E}\left[|D(k,l)|^2\right]$.

3. PROPOSED ALGORITHM

The proposed algorithm for transient interference suppression is divided into two steps. First, in Subsection 3.1, we propose a modified version of the OM-LSA adapted to track fast variations and used for transient PSD estimation. Then, in Subsection 3.2, we compute an OM-LSA filter using the PSD estimate from Subsection 3.1 to enhance the speech. Figure 1 depicts a block diagram of the proposed algorithm.

3.1. Transient PSD Estimation

We exploit the different variation rates of speech, transients and background noise. We assume that the transient interference is rapidly varying in time compared to the slower speech and the pseudo-stationary background noise. Thus, we propose to adjust the noise PSD estimation component of the OM-LSA to track faster PSD changes and to make the non-transient components (both speech and background noise) appear as "pseudo-stationary". Then, computing the OM-LSA estimator based on the PSD estimate of the non-transient components enables to enhance the transient part and suppress the speech and background noise.

We use short STFT frames which reduce the speech variation between consecutive time frames. We note that the usage of short time frames reduces the frequency resolution. In our empirical experiments, a time frame length of 64 samples was shown to be suitable and yielded good performance. This particular length corresponds to 4 ms for a 16 KHz sampling rate.

We modify the *minima controlled recursive averaging* (MCRA) method [9] to estimate the PSD of the non-transient components. The PSD estimation is based on a smoothed periodogram obtained by a temporal recursive averaging of the spectral amplitude as follows

$$S(k,l) = \alpha_s S(k,l-1) + (1-\alpha_s)|Y(k,l)|^2. \quad (5)$$

Thus, assigning a smaller value to the smoothing parameter α_s enables faster tracking of the PSD. The lower α_s is, the more weight is assigned to the current time frame, and as a result faster variations of the PSD of the speech or background noise can be captured. In the proposed algorithm we reduce the value of α_s from a typical range of $0.9 - 0.99$ to 0.7.

The transient presence probability is controlled by the minima values of the smoothed periodogram [10], which are obtained from a finite causal window of length L

$$S^L_{min}(k,l) = \min\{S(k,l), S(k,l-1)\ldots S(k,l-L+1)\}. \quad (6)$$

Then, the transient presence decision is made using the following rule

$$S_r(k,l) \equiv \frac{S(k,l)}{S^L_{min}(k,l)} \lessgtr \delta \quad (7)$$

where δ is an empirical threshold. When $S_r(k,l) > \delta$ the current slot is marked as containing a transient. Otherwise, it is regarded as containing speech and background noise. Let $I(k,l)$ denote the transient presence indicator, defined as

$$I(k,l) = \begin{cases} 1, & \text{if } S_r(k,l) > \delta, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

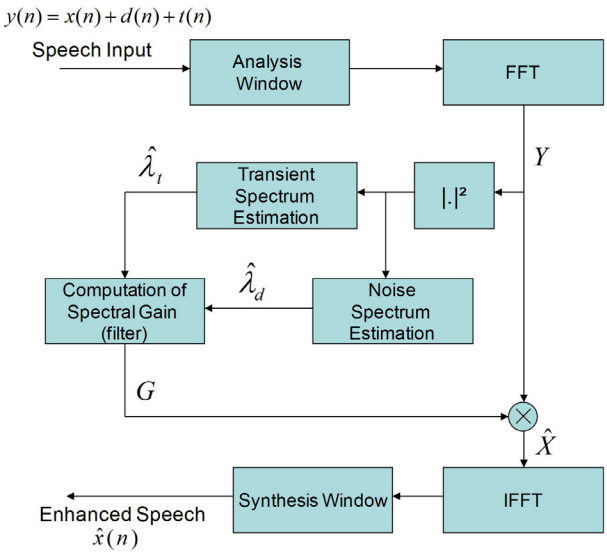

Fig. 1. Block diagram of the proposed algorithm.

Let $p(k,l)$ be the transient presence probability, which is smoothed according to

$$p(k,l) = \alpha_p p(k,l-1) + (1-\alpha_p)I(k,l) \quad (9)$$

where α_p $(0 < \alpha_p < 1)$ is a smoothing parameter. Finally, the PSD of the non-transient components is estimated by averaging past spectral power values using a smoothing parameter that is adjusted by the transient presence probability:

$$\widehat{\lambda}(k,l+1) = \widetilde{\alpha}(k,l)\widehat{\lambda}(k,l) + [1-\widetilde{\alpha}(k,l)]|Y(k,l)|^2 \quad (10)$$

where

$$\widetilde{\alpha}(k,l) \doteq \alpha + (1-\alpha)p(k,l) \quad (11)$$

and α is a fixed smoothing parameter $(0 < \alpha < 1)$.

The proposed procedure captures most of the speech and background noise parts. Unfortunately, speech phoneme onsets, which are characterized by sudden bursts, are not tracked by the spectral recursive smoothing. Therefore, according to (7), speech phoneme onsets may be wrongly considered as transients and as a result may not be properly suppressed in this stage. We propose to take into account "future" time frames in order to distinguish between transients and speech onsets when a sudden burst is encountered. After a short-duration transient, the power of the signal is expected to decay fast, whereas after a speech phoneme onset, the power level is expected to stay steady for the duration of the phoneme.

We implement this distinction using an additional anti-causal window of length T. Let $S^T_{min-ac}(k,l)$ be the minimum value in the anti-causal window, which is calculated according to

$$S^T_{min-ac}(k,l) = \min\{S(k,l), S(k,l+1)\ldots S(k,l+T-1)\}. \quad (12)$$

Then, the maximal value of the two minimal spectra values from the causal and anti-causal windows is computed

$$S_{min}(k,l) = \max\{S^L_{min}(k,l), S^T_{min-ac}(k,l)\}, \quad (13)$$

and regarded as the estimate of the non-transient spectral level. This ensures that onsets spectra are compared to the rest of the

phoneme rather than the preceding background noise. By substituting $S^L_{min}(k,l)$ with $S_{min}(k,l)$ in (7) we obtain the new decision rule. Since the phoneme onset spectral level is now not higher than S_{min}, the value of S_r does not cross the threshold δ and the onsets are suppressed as desired.

We observe that an anti-causal window shorter than a typical transient yields unwanted tracking of transients, whereas a window longer than a typical speech phoneme captures the spectral minimal level of the stationary noise instead of the speech phoneme. Thus, the length T of the anti-causal window is set to be longer than a typical transient and shorter than a typical speech phoneme. Our empirical tests suggest that a 40ms long anti-causal window is appropriate for many transients. However, the window length should generally be set according to the specific transient interferences. We also note that an inevitable consequence of using an anti-causal window is the addition of time lag.

Finally, we use the modified OM-LSA output signal, which contains mainly the energy of the transient, to estimate the PSD of the transient as follows

$$\widehat{\lambda}_t(k,l) = |\widehat{T}(k,l)|^2 \tag{14}$$

where $\widehat{T}(k,l)$ is the modified OM-LSA output transient amplitude estimation. This estimation is performed for the simplicity of the discussion. Alternatively, the PSD of the transient interference could be estimated directly based on the a-priory SNR [4].

3.2. Speech Enhancement

In this section, we propose to use the output of Subsection 3.1 (the transient PSD estimate $\widehat{\lambda}_t(k,l)$) as an additional input for a second application of the OM-LSA filter as presented in Figure 1.

As described in [4], the OM-LSA algorithm assembles an optimal log spectral amplitude (LSA) filter in which the transient PSD estimate is incorporated into the filter computation and enables transient interference suppression. We compute a new total interference PSD estimate, which is given by

$$\widehat{\lambda}^*_d(k,l) = \widehat{\lambda}_d(k,l) + \widehat{\lambda}_t(k,l) \tag{15}$$

where $\widehat{\lambda}_d$ is the estimate of the stationary noise PSD obtained by the original MCRA, and $\widehat{\lambda}_t$ is the estimate of the transient noise PSD from Subsection 3.1. Let $\xi \doteq \frac{\lambda_x(k,l)}{\lambda^*_d(k,l)}$ and $\gamma \doteq \frac{|Y(k,l)|^2}{\lambda^*_d(k,l)}$ be the a-priori and posteriori SNRs, respectively. Then, the spectral gain is derived according to

$$G(k,l) = \{G_{H_1}(k,l)\}^{p(k,l)} G_{min}^{1-p(k,l)} \tag{16}$$

where G_{min} is a constant low gain used when speech is absent, and

$$G_{H_1}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left(\frac{1}{2}\int_{v(k,l)}^{\infty} \frac{e^{-t}}{t} dt\right), \tag{17}$$

with

$$v(k,l) = \frac{\gamma(k,l)\xi(k,l)}{1-\xi(k,l)}.$$

As shown in [4], the optimal gain $G(k,l)$ minimizes the mean-square error of the LSA under speech presence uncertainty

$$\min\{(\log(A(k,l)) - \log(\hat{A}(k,l)))\}$$

where $A(k,l) = |X(k,l)|$ is the speech spectral amplitude.

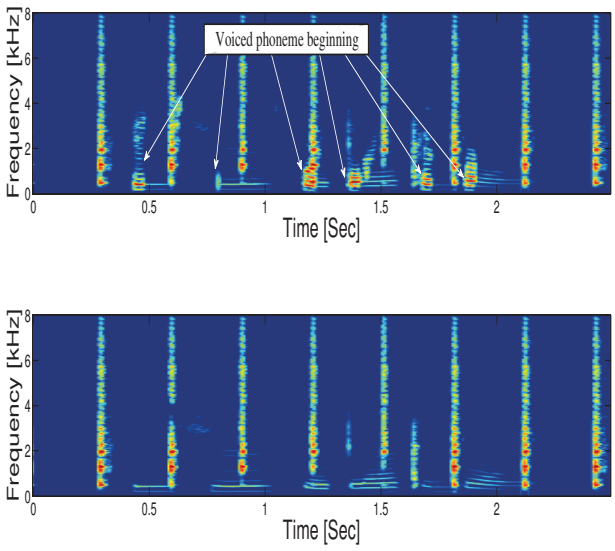

Fig. 2. Estimated transient interference PSD. Top: *without* the anti-causal window. Bottom: *with* the anti-causal window.

4. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the proposed algorithm. We use recorded speech signals of different speakers, both male and female, with an average length of 3 s. We arbitrarily add $4-15$ recorded transient instances of $40-140$ ms durations along the speech signal. The transients and speech signals are taken from [11] and [12], respectively, and are both sampled at 16 KHz. We note that unlike [6, 5, 7] the proposed method does not require several repetitions of the transient instances. For the first stage of the transient PSD estimation we use STFT frames of length 64 and for the speech enhancement stage we use longer frames of length 512. In both cases, we use 75% overlap between consecutive frames. The amplitudes of the speech and the transients are re-scaled to the same maximal amplitude value. This provides a fair comparison between different types of speech and transients. We employ the algorithm on signals contaminated with 5 types of transients.

Figure 2 demonstrates the onsets problem described in Section 3.1 by presenting the output of the first stage of the algorithm. In Fig. 2 (top) we present the PSD estimate without the anti-causal window. We observe that the speech onsets are recognized as transients and therefore are not suppressed. In Fig. 2 (bottom) we present the PSD estimate with the anti-causal window. As shown, the beginnings of the phonemes are suppressed while transient interferences remain undistorted.

Figure 3 (top) depicts the spectrograms of a speech signal contaminated by a metronome interference, and Fig. 3 (bottom) depicts the enhanced speech. It can be seen that the transients are of short duration in time and span a wide range of frequencies. In Fig. 3 (bottom) we observe significant suppression of the interference while imposing merely a small distortion on the speech.

The results are evaluated using a common objective measure - the Segmental SNR (SegSNR) and summarized in Table 1. It can be seen clearly that the proposed method enables to suppress a variety of transient interferences and improves substantially the SegSNR of

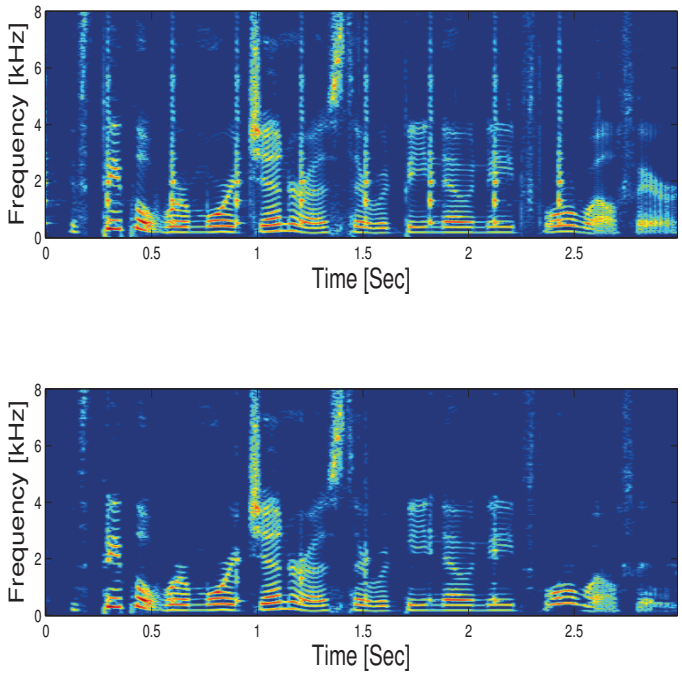

Fig. 3. Signal spectrograms. Top: the noisy measurement. Bottom: the enhanced speech obtain by the proposed algorithm.

Table 1. Speech enhancement evaluation for different types of transient interference.

Transient Noise	SegSNR Improvement [dB]	delay [ms]
Metronome	4.7	40
Knocks	4.9	40
Shot	3.7	250
Scissors	5.3	100
Keyboard	4.5	40

the signal. We mention that the length of the anti-causal window is empirically set for each type of transient to yield maximal performance; The obtained SegSNR for each type of transient is measured for different lengths of anti-causal windows. Then, the length that yields the maximal SegSNR is set. The determined lengths of the anti-causal windows are presented in Table 1. The obtained results demonstrate the robustness of the proposed method. Transient interferences of various types, durations, and spectral features are reduced regardless of their location along the speech and regardless of the particular speaker.

5. CONCLUSIONS

We introduced an algorithm for transient interference suppression based on two applications of the OM-LSA estimator. First, the transient interference is enhanced and its PSD is estimated using a version of the OM-LSA adjusted to track rapid signal variations. In addition, in this step we use both causal and anti-causal windows for minimum statistics tracking. This circumvents false identification of voiced phonemes onsets as transients. Then, the pseudo-stationary noise PSD is estimated using the MCRA method. The sum of the PSDs of the transient interfernce and pseudo-stationary noise is utilized as the total interference PSD estimate. Finally, the total interference PSD estimate is used in the OM-LSA filter for speech enhancement.

The proposed algorithm enables to suppress a variety of transients. In addition, the algorithm is shown to be speaker independent. A particularly good performance is achieved for transients that span a wide frequency range and have short temporal support. Experimental results demonstrate successful suppression of transients without a-priori knowledge on their location along the speech signal. Moreover, unlike previous methods, several transient instances are not required for successful suppression and even a single transient can be suppressed with minimal speech distortion.

6. REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech, and Signal Process.*, vol. 27, no. 2, pp. 113 – 120, Apr. 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancememt using a minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Acoust. Speech and Signal Process.*, pp. 1109–1121, Dec. 1984.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log spectral amplitude estimator," *IEEE Trans. Acoust. Speech and Signal Process.*, pp. 443–445, Apr. 1985.

[4] I Cohen and B. Berdugo, "Speech enhancement for non stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.

[5] R. Talmon, I. Cohen, and S. Gannot, "Transient noise reduction using nonlocal diffusion filters," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, Issue 6, pp. 1584–1599, Aug. 2011.

[6] R. Talmon, I. Cohen, and S. Gannot, "Clustering and suppression of transient noise in speech signals using diffusion maps," *Proc. 36th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP11), Prague, Czech Republic*, May 22-28, 2011.

[7] R. Talmon, I. Cohen, and S. Gannot, "Single-channel transient interference suppression with diffusion maps," *to appear in IEEE Trans. Audio, Speech and Lang. Process.*, Apr. 2012.

[8] R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, pp. 5–30, Jul. 2006.

[9] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[10] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[11] *[Online]. Available: http://www.freesound.org*.

[12] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic-phonetic continous speech database," National Inst. of Standards and Technology (NIST), Gaithersburg, MD, Feb 1993.